Argonne NATIONAL LABORATORY

... for a brighter future

ALCF

Argonne Leadership Computing Facility

Overview and Blue Gene/L Hardware Architecture

Argonne Leadership Computing Facility

Ray Bair, Director
Argonne National Laboratory and University of Chicago

February 7, 2007

U.S. Department of Energy

UChicago ▶ Argonne LLC

Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Argonne Leadership Computing Facility

- What do they do at Argonne?
- What's Leadership Computing?
- Where did ALCF come from?
- Where is it going?

About Argonne

- Founded in 1943, designated a national laboratory in 1946
- Managed by The University of Chicago for the U.S. Department of Energy
  - More than 2,900 employees and 5,000+ facility users
  - About $475M budget
  - 1,500-acre, wooded site in DuPage County, Illinois
- Broad R&D portfolio
- Numerous sponsors

Argonne's Mission

- Serve DOE & national security
  - Advancing the frontiers of knowledge
  - Creating and operating forefront scientific user facilities (e.g., Advanced Photon Source, Intense Pulsed Neutron Source, Argonne Tandem-Linac Accelerator System)
  - Providing innovative and effective tools and solutions for energy and environmental challenges to national and global well-being, in the near and long term
- In accomplishing its mission, Argonne partners with DOE, other federal labs, academia, and the private sector

Forefront Science and Engineering

- Basic and applied research
  - Materials and chemical sciences and engineering
  - High energy, nuclear, and atomic physics
  - Multidisciplinary nanoscience and nanotechnology
  - Structural biology, functional genomics, and bioinformatics
  - Environmental science, technology, and assessment
  - Transportation technology
  - Computer science and applied mathematics
  - Computational science
- Design, construction, and operation of accelerator-based user facilities
- Design, development, and evaluation of advanced nuclear energy systems and proliferation-resistant nuclear fuel-cycle technologies

DOE Leadership Computing Facility Strategy

- DOE-SC selected the ORNL, ANL and PNNL team (May 12, 2004), based on a competitive peer review of 4 LCF proposals
  - ORNL will deploy a series of systems based on Cray's XT3/4 architectures @ 250TF/s in FY07 and 1000TF/s in FY08/9
  - ANL will develop a series of systems based on IBM's BlueGene @ 100TF/s in FY07 and 250-500TF/s in FY08/FY09 with IBM Blue Gene/P
  - PNNL will contribute software technology
- DOE SC will make these systems available as capability platforms to the broad national community via competitive awards (e.g., INCITE Allocations)
  - Each facility will target ~20 large-scale production applications teams
  - Each facility will also support development users
- DOE's LCFs complement existing and planned production resources at NERSC
  - Capability runs will be migrated to the LCFs, improving NERSC throughput
  - NERSC will play an important role in training and new user identification

Over 20 years of Advanced Systems for DOE and Others

- **ACRF** period [1983-1992]
  - DOE's founding ACRF
  - Explored many parallel architectures, developed programming models and tools, trained >1000 people
- **HPCRC** period [1992-1999]
  - Production-oriented parallel computing for Grand Challenges in addition to Computer Science.
  - Fielded 1st IBM SP in DOE

- **TeraGrid** [2001-present]
  - Overall Project Lead
  - Defining, deploying and operating the integrated national cyberinfrastructure for NSF
  - 9 sites, 22 systems, 200TF
- **LCRC** [2003-present]
  - Lab-wide production supercomputer service
  - All research divisions, 56 projects, 380 users
- **BlueGene Evaluation** [2005-present]
  - Founded BlueGene Consortium with IBM
    - *67 institutions, >260 members*
    - *Applications Workshop Series*
    - *Systems Software Collaborations*

The Blue Gene Consortium

formed by Argonne and IBM, April 2004

- Focuses interest in the Blue Gene series
  - Exploiting its potential for computational science
- Creates a framework for cooperation
  - Developing applications, tools and systems software
  - Sharing operations and support strategies
  - Exchanging innovations and novel solutions
- Supports upcoming HPC needs
  - Access to Argonne BG/L for evaluation
  - Blue Gene Watson days for 8-20 rack runs
  - Community participation in requirements for next generation systems

Working Groups
- **Applications**
- **System Software**
- **Operations**
- **Architecture**
- **Outreach**

http://www.mcs.anl.gov/bgconsortium/

Blue Gene Consortium Members (67)

DOE Laboratories

- Ames National Laboratory/Iowa State U.
- Argonne National Laboratory
- Brookhaven National Laboratory
- Fermi National Laboratory
- Jefferson Laboratory
- Lawrence Berkeley National Laboratory
- Lawrence Livermore National Laboratory
- Oak Ridge National Laboratory
- Pacific Northwest National Laboratory
- Princeton Plasma Physics Laboratory

Universities

- Boston University
- California Institute of Technology
- Columbia University
- Cornell University
- DePaul University
- Harvard University
- Illinois Institute of Technology
- Indiana University
- Iowa State University
- Louisiana State University
- Massachusetts Institute of Technology
- National Center for Atmospheric Research
- New York University/Courant Institute
- Northern Illinois University

Universities (continued)

- Northwestern University
- Ohio State University
- Pennsylvania State University
- Pittsburgh Supercomputing Center
- Princeton University
- Purdue University
- Rutgers University
- Stony Brook University
- Texas A&M University
- University of California – Irvine
- University of California – San Diego/SDSC
- University of California – San Francisco
- University of Chicago
- University of Colorado – JILA
- University of Delaware
- University of Hawaii
- University of Illinois – Urbana Champaign
- University of Minnesota
- University of North Carolina
- University of Southern California/ISI
- University of Tennessee
- University of Texas at Austin – TACC
- University of Utah
- University of Wisconsin

Industry

Engineered Intelligence Corporation
Gene Network Services
IBM
Raytheon
Visual Numerics Inc.

International

Allied Engineering Corp., Japan
ASTRON/LOFAR, The Netherlands
Centre of Excellence for Applied Research and Training, UAE
Ecole Polytechnique Fédérale de Lausanne, Switzerland
Institut de Physique de Globe de Paris, France
KTH - Royal Institute of Technology,
National Institute of Advanced Industrial Science & Tech., Japan
National University of Ireland
Trinity College, Ireland
John von Neumann Institute, Germany
NIWS Co., Ltd., Japan
University of Edinburgh, EPCC Scotland
University of Paris, France
University of Tokyo, Japan

www.mcs.anl.gov/bgconsortium

26 BlueGene/L Systems on 11/06 TOP500 List

819 teraflops fielded today, with 294,912 processors

Rank	Site	Country	Processors	RMax	RPeak
1	DOE/NNSA/LLNL	United States	131,072	280,600	367,000
3	IBM Thomas J. Watson Research Center	United States	40,960	91,290	114,688
13	Forschungszentrum Juelich (FZJ)	Germany	16,384	37,330	45,875
17	ASTRON/University Groningen	Netherlands	12,288	27,450	34,406
21	Computational Biology Research Center, AIST	Japan	8,192	18,200	22,938
22	Ecole Polytechnique Federale de Lausanne	Switzerland	8,192	18,200	22,938
23	High Energy Accelerator Research Org (KEK)	Japan	8,192	18,200	22,938
24	High Energy Accelerator Research Org (KEK)	Japan	8,192	18,200	22,938
25	IBM - Rochester Deep Computing	United States	8,192	18,200	22,938
42	UCSD/San Diego Supercomputer Center	United States	6,144	13,780	17,203
52	IBM - Rochester DD1 Prototype	United States	8,192	11,680	16,384
63	High Energy Accelerator Research Org (KEK)	JAPAN	4,096	9,360	11,469
64	IBM - Almaden Research Center	United States	4,096	9,360	11,469
65	IBM Research	Switzerland	4,096	9,360	11,469
66	IBM Thomas J. Watson Research Center	United States	4,096	9,360	11,469
138	Argonne National Laboratory	United States	2,048	4,713	5,734
139	Boston University	United States	2,048	4,713	5,734
140	CERT (Center for Excellence for Applied R&T)	UAR	2,048	4,713	5,734
141	Iowa State University	United States	2,048	4,713	5,734
142	Lawrence Livermore National Laboatory	United States	2,048	4,713	5,734
143	MIT	United States	2,048	4,713	5,734
144	NCAR (National Center for Atmospheric Research)	United States	2,048	4,713	5,734
145	NIWS Co, Ltd	Japan	2,048	4,713	5,734
146	Princeton University	United States	2,048	4,713	5,734
147	Renaissance Computing Institute (RENCI)	United States	2,048	4,713	5,734
148	University of Edinburgh	United Kingdom	2,048	4,713	5,734

Mission and Vision for the ALCF

Our Mission

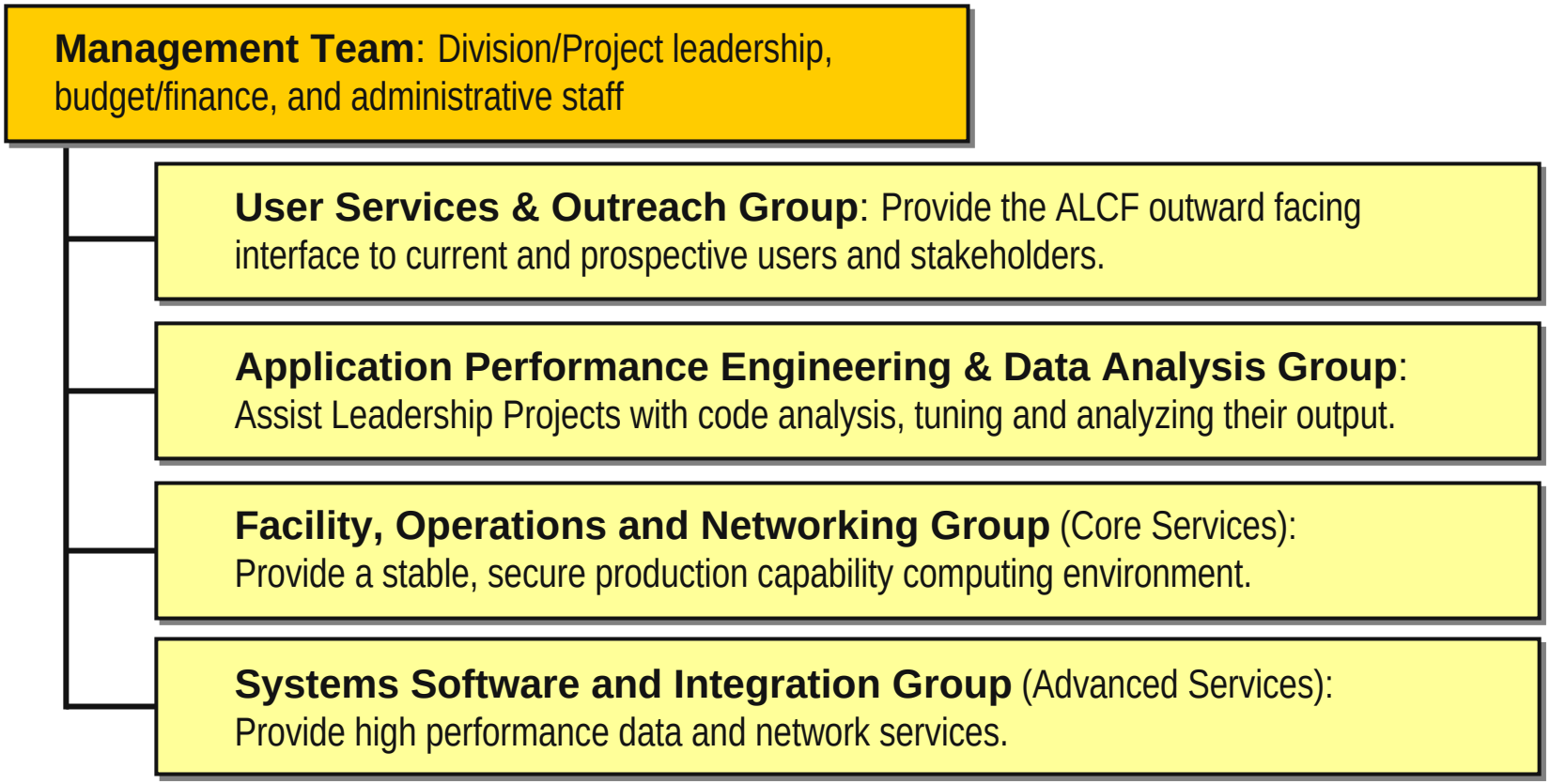
Provide the computational science community with a world leading computing capability dedicated to breakthrough science and engineering.

Our Vision

A world center for computation driven scientific discovery that has:

- outstandingly talented people,
- the best collaborations with computer science and applied mathematics,
- the most capable and interesting computers, and
- a true spirit of adventure.

ALCF Organization



ALCF Timeline

2004

- Formed the Blue Gene Consortium with IBM

2005

- Installed 5 teraflops Blue Gene/L for evaluation

2006

- Began production support of 6 INCITE projects
- Continued code development and evaluation

2007

- Increase to 9 INCITE projects; continue development projects
- Blue Gene/P workshop (May-June)
- Install 100 teraflops Blue Gene/P system (Oct.-Nov.)

2008

- Begin support of INCITE projects on Blue Gene/P
- Add 250-500 teraflops Blue Gene/P system

2007 INCITE Projects on Blue Gene/L

10M Hours on BGL at ALCF + BGW at IBM Watson Res. Ctr.

New Projects

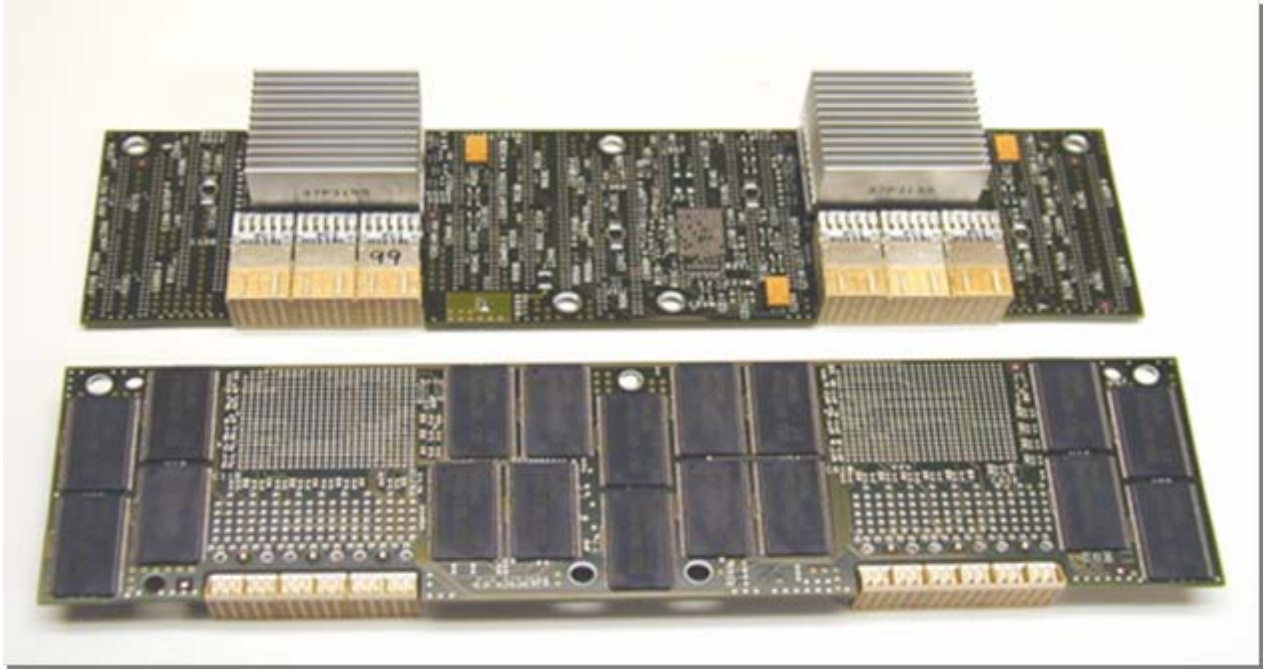
- Argonne National Laboratory: Reactor Core Hydrodynamics
- Oak Ridge National Laboratory: Statistical Physics of Fracture
- Northwestern University: Coherent Control of Light in Nanoscale
- Procter and Gamble Co.: Molecular Simulations of Surfactant-Assisted Aqueous Foam Formation
- University of California–Davis: Water in Confined States

Renewals

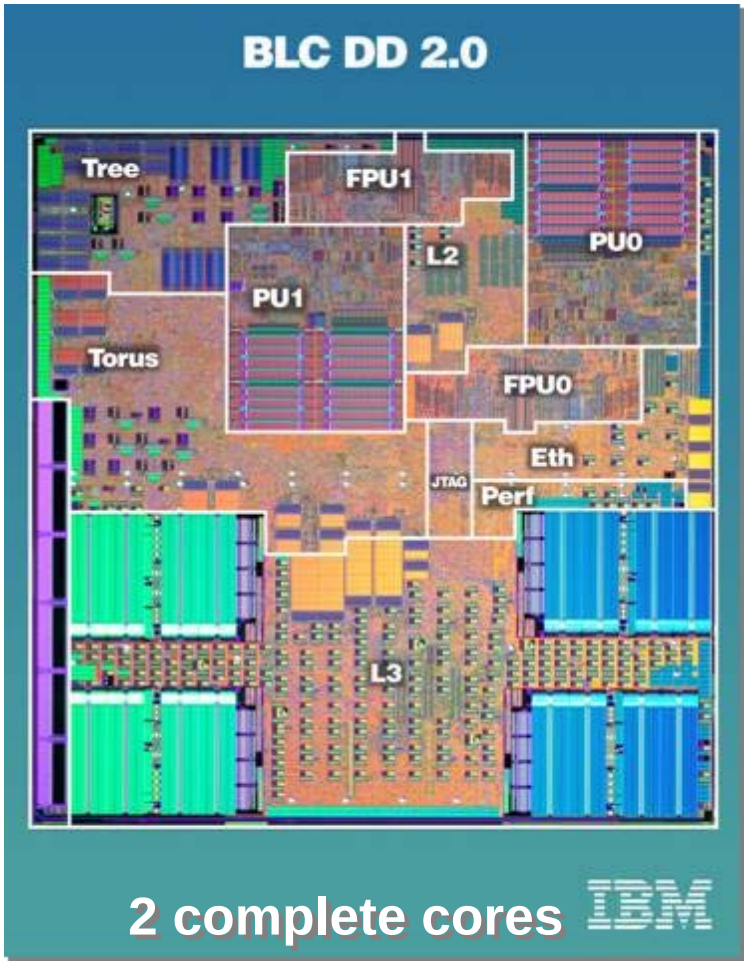
- Pratt and Whitney: High-Fidelity LES Simulations of an Aircraft Engine Combustor to Improve Emissions and Operability
- University of Alaska–Fairbanks: Modeling the Response of Terrestrial Ecosystems to Climate Change and Disturbance
- University of California–San Diego: Simulation and Modeling of Synuclein-based "Protofibril Structures" as a Means of Understanding the Molecular Basis of Parkinson's Disease
- University of Washington: High-Resolution Protein Structure Prediction

Overview of the Blue Gene/L System Architecture

- What's in the black box?
- How is it different than other computers?
- What does the hardware look like to applications?



BlueGene/L Chip



Just add DRAM

Source: IBM

Processor

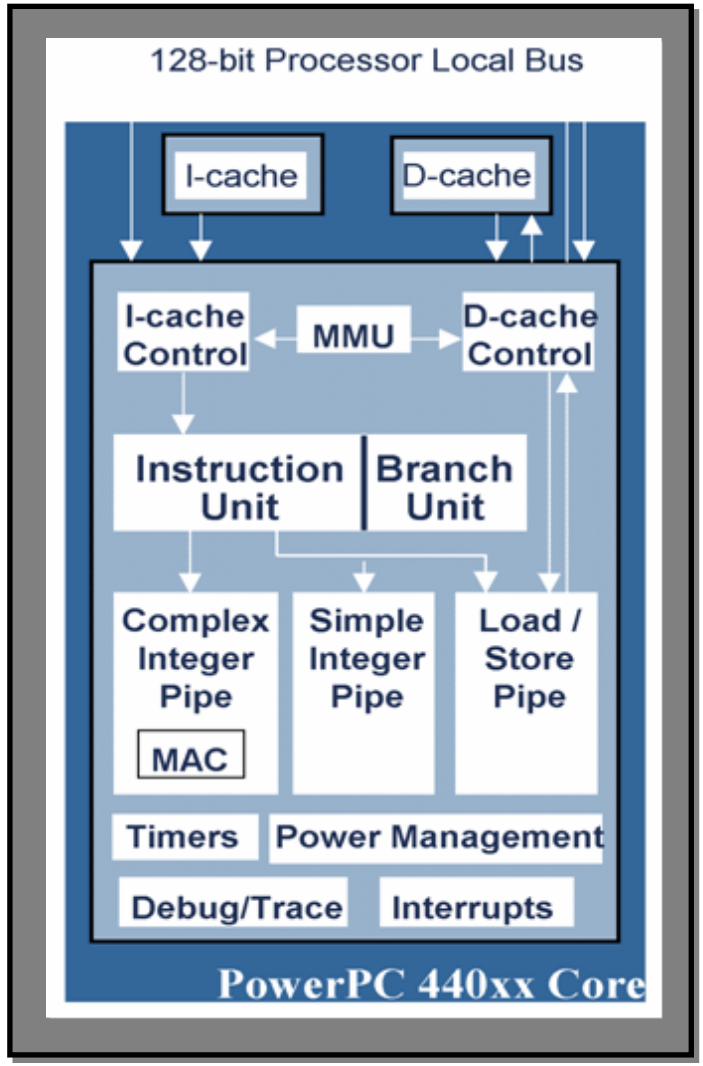
- PPC440x5 Processor Core – 700 MHz
  - *Superscalar: 2 instructions per cycle*
  - *Out of order issue and execution*
  - *Dynamic branch prediction, etc.*
- Two 64-bit floating point units
  - *SIMD instruct. over both register files*
  - *Parallel (quadword) loads/stores*
  - *2.8 GFLOPS/processor*

Interconnect

- 3-Dimensional Torus
  - *Virtual cut-through hardware routing*
  - *1.4Gb/s on all 12 node links*
  - *1 µs latency bet. neighbors, 5 µs to farthest*
- Global Tree
  - *One-to-all broadcast, reduction functionality*
  - *2.8 Gb/s of bandwidth per link*
  - *Latency of one-way tree traversal 2.5 µs*
- Low Latency Global Barrier and Interrupt
  - *Latency of round trip 1.3 µs*
- Ethernet
  - *All external comm. (file I/O, control, etc.)*
- Control Network

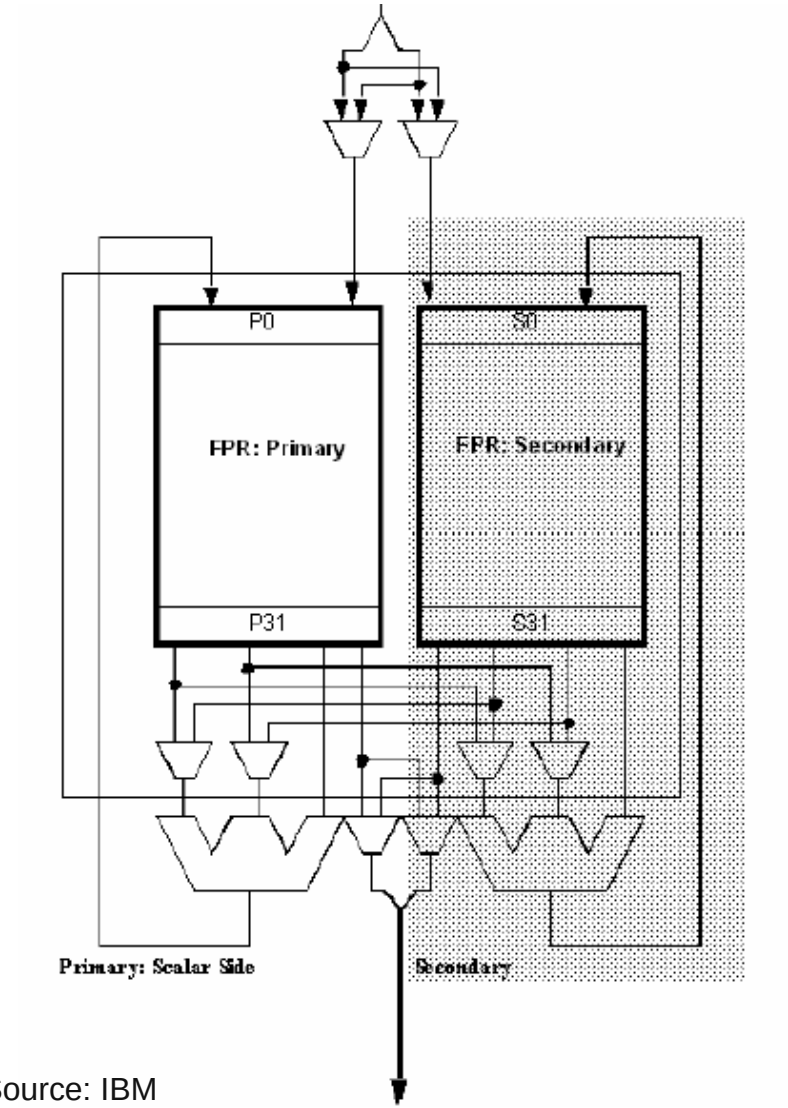
PPC440x5 Processor Core Features

- High-performance embedded PowerPC core
- 2.0 DMIPS/MHz
- Book E Architecture
- Superscalar: Two instructions per cycle
- Out of order issue, execution, and completion
- 7 stage pipeline
- 3 Execution pipelines
  - Combined complex, integer, & branch pipeline
  - Simple integer pipeline
  - Load/store pipeline
- Dynamic branch prediction
- Single cycle multiply
- Single cycle multiply-accumulate
- Real-time non-invasive trace
- 128-bit CoreConnect Interface



Source: IBM

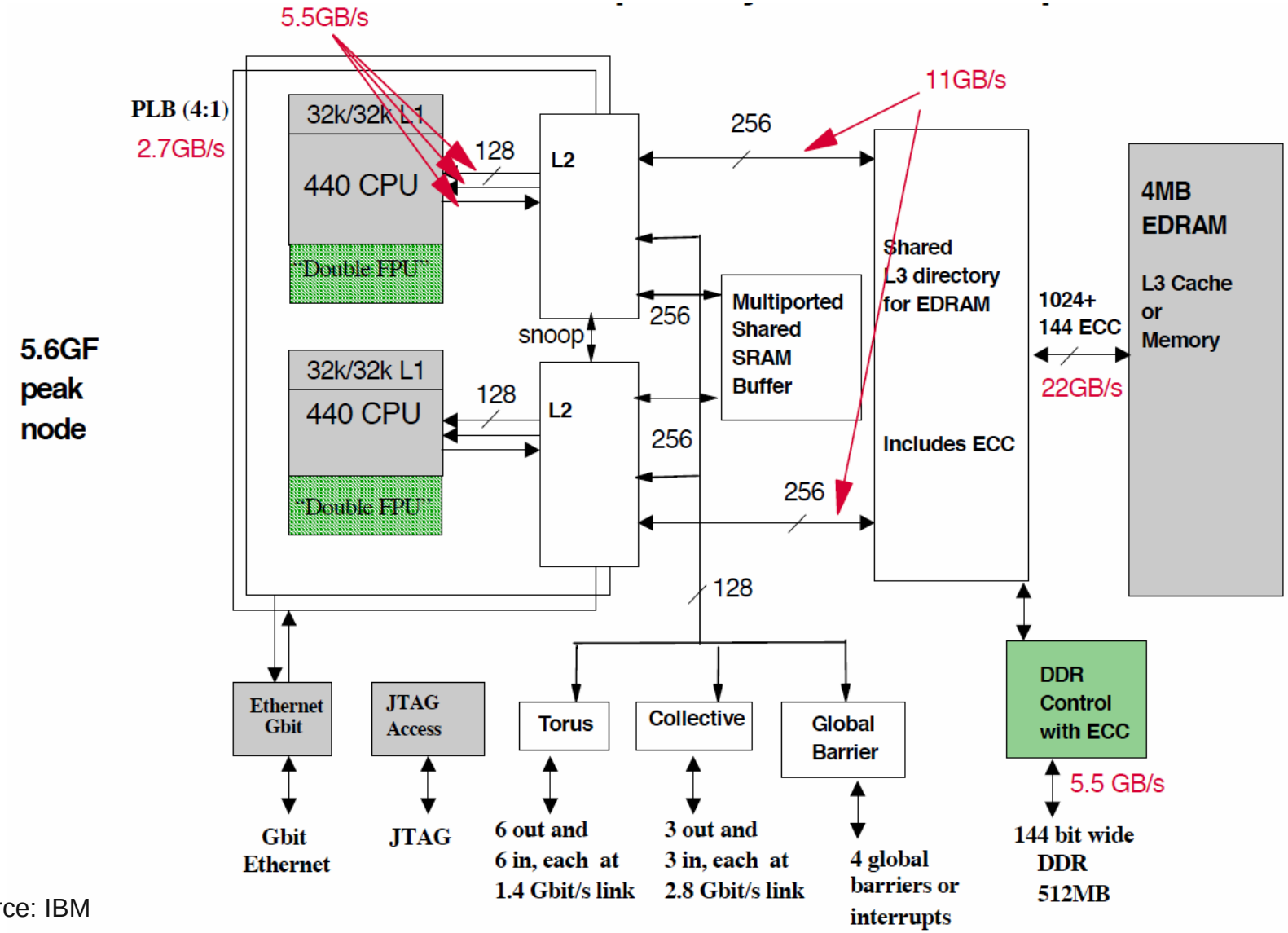
Dual FPU Architecture



Source: IBM

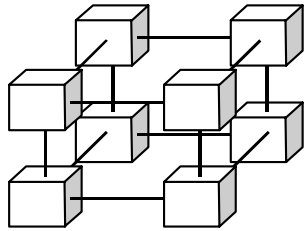
- Two 64 bit floating point units
- Designed with input from compiler and library developers
- SIMD instructions over both register files
  - FMA operations over double precision data
  - More general operations available with cross and replicated operands
    - *Useful for complex arithmetic, matrix multiply, FFT*
- Parallel (quadword) loads/stores
  - Fastest way to transfer data between processors and memory
  - Data needs to be 16-byte aligned
  - Load/store with swap order available
    - *Useful for matrix transpose*

BlueGene/L Compute System on a Chip ASIC

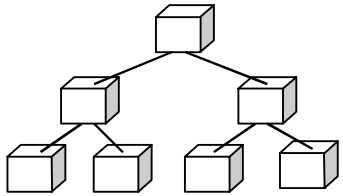


Source: IBM

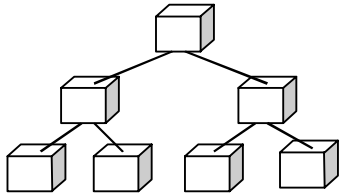
BlueGene/L – Five Independent Networks



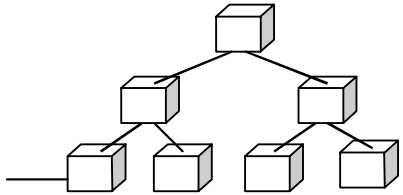
3-Dimensional Torus
- Point-to-point



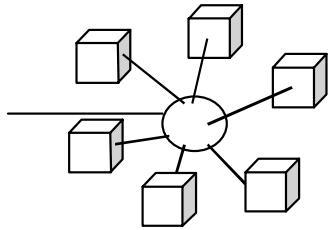
Global Tree
- Global Operations



Global Barriers and Interrupts
- Low Latency Barriers and Interrupts



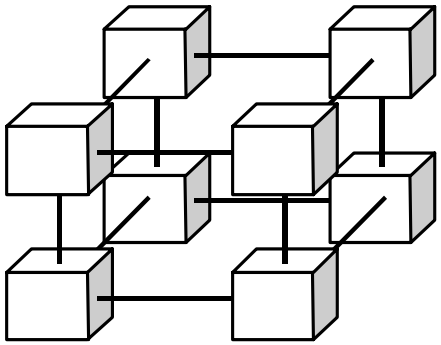
Gbit Ethernet
- File I/O and Host Interface

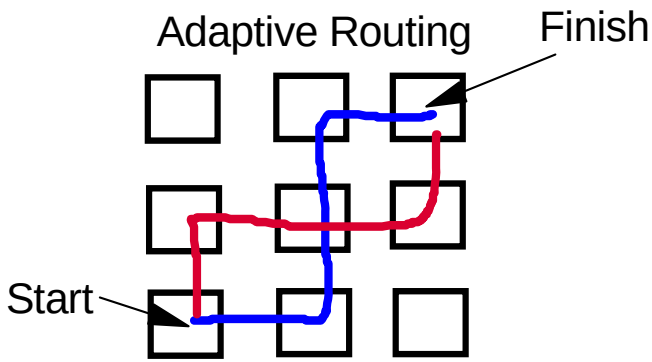


Control Network
- Boot, Monitoring and Diagnostic

Source: IBM

3-D Torus Network



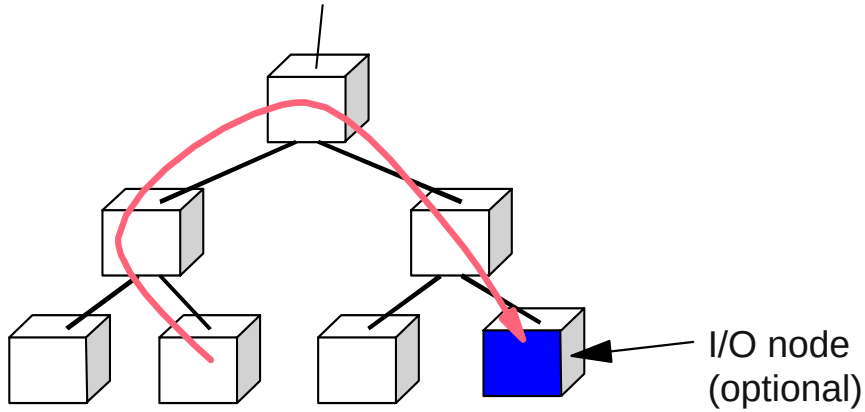


- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * * 1.4Gb/s = 68 TB/s total Torus bandwidth
- 4 *32 *32 * 1.4Gb/s = 5.6 Tb/s bi-sectional bandwidth
- Worst case hardware latency through node ~ 69nsec

- Virtual cut-through routing with multipacket buffering on collision
  - Minimal
  - Adaptive
  - Deadlock Free
- Class Routing Capability (Deadlock-free Hardware Multicast)
  - Packets can be deposited along route to specified destination.
  - Allows for efficient one to many in some instances.
- Active messages allow for fast transposes as required in FFTs.
- Independent on-chip network interfaces enable concurrent access.

Source: IBM

Tree Network

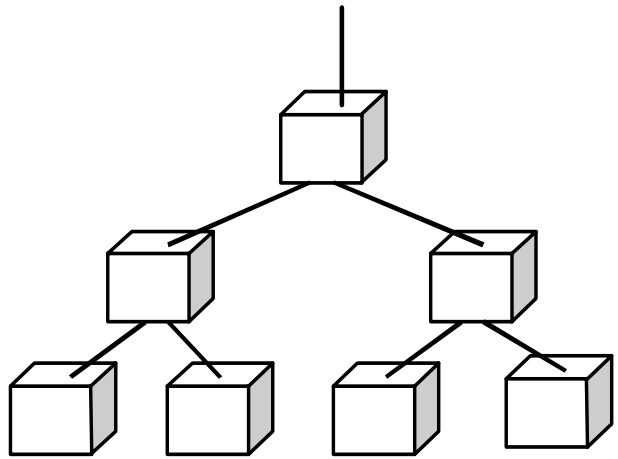


- High Bandwidth one-to-all
  - 2.8Gb/s to all 64k nodes
  - 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5μsec to top, additional 2.5μsec to broadcast to all
- Global sum over 64K in less than 2.5μsec (to top of tree)
- Used for disk/host funnel in/out of I/O nodes
- Minimal impact on cabling
- Partitioned with Torus boundaries
- Flexible local routing table
- Used as Point-to point for File I/O and Host communications

Source: IBM

Fast Barrier Network

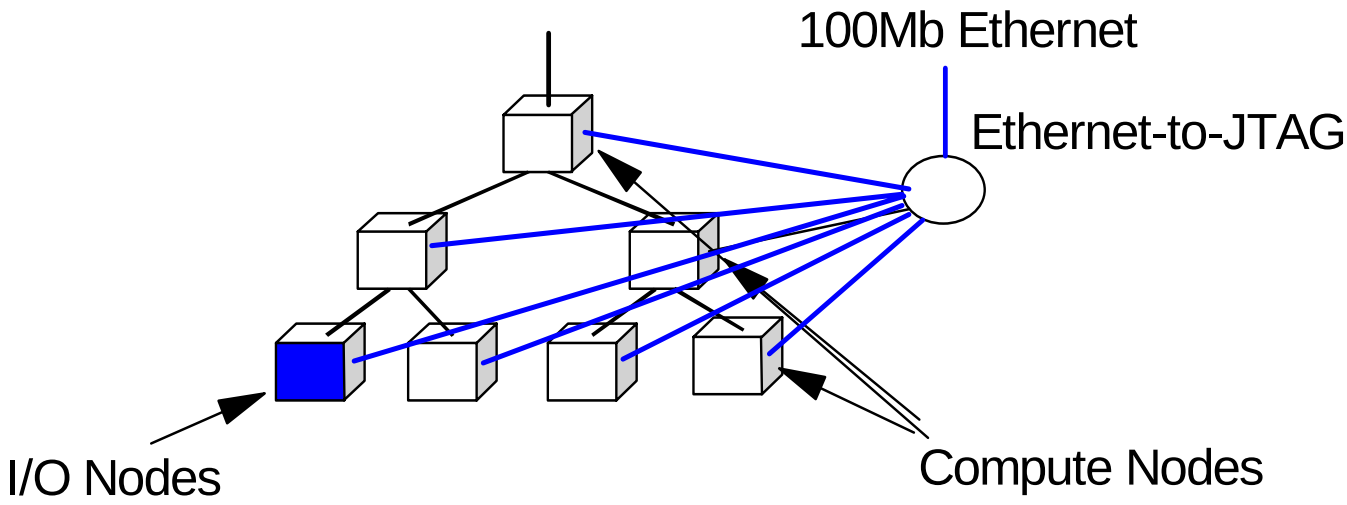
- Four Independent Barrier or Interrupt Channels
  - Independently Configurable as “or” or “and”
- Asynchronous Propagation
  - Halt operation quickly
    (current estimate is 1.3μsec worst case round trip)
    > 3/4 of this delay is time-of-flight
- Sticky bit operation
  - Allows global barriers with a single channel
- User Space Accessible
  - System selectable
- Partitions along same boundaries as Tree and Torus
  - Each user partition contains its own set of barrier/interrupt signals.



Source: IBM

Control Network

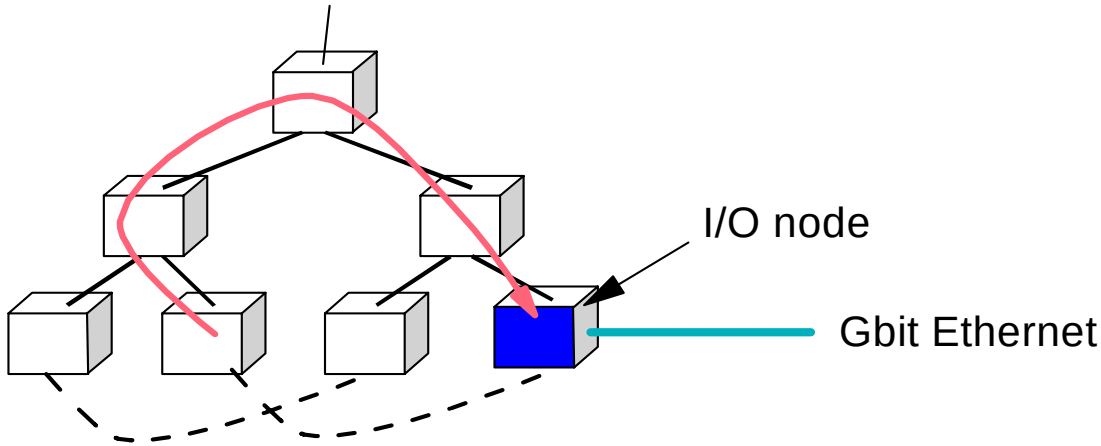
- JTAG interface to 100Mb Ethernet
  - Direct access to all nodes
  - Boot, system debug availability
  - Runtime noninvasive RAS support
  - Noninvasive access to performance counters
  - Direct access to shared SRAM in every node



Source: IBM

Ethernet Disk/Host I/O Network

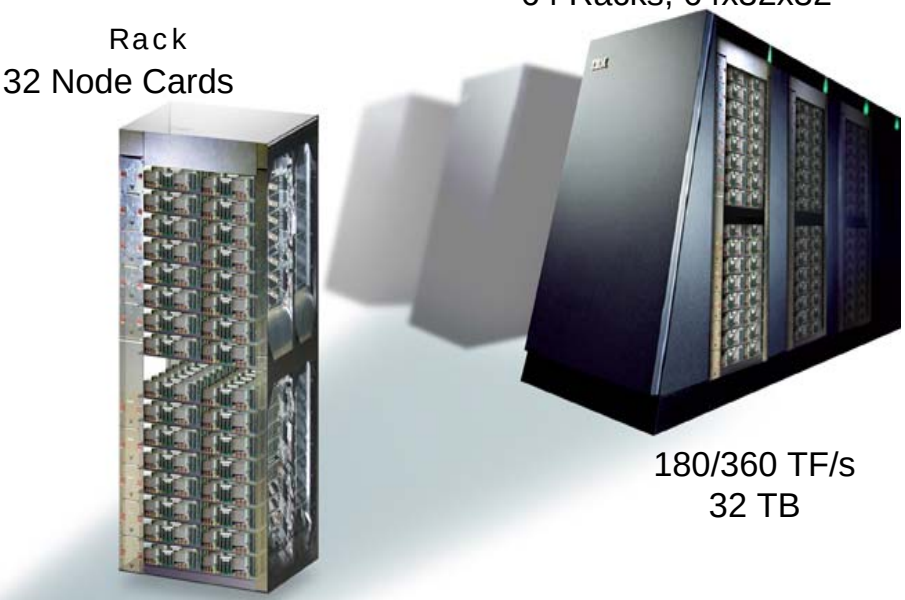
- Gb Ethernet on all I/O nodes
  - Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.
  - Funnel via global tree.
  - I/O nodes use same ASIC but are dedicated to I/O Tasks.
  - I/O nodes can utilize larger memory.
- Dedicated DMA controller for transfer to/from Memory
- Configurable ratio of Compute to I/O nodes
  - I/O nodes are leaves on the tree network.



Source: IBM

The Blue Gene Family of Computers

- Puts processors + memory + network interfaces on same chip.
- Achieves good compute-communications balance.

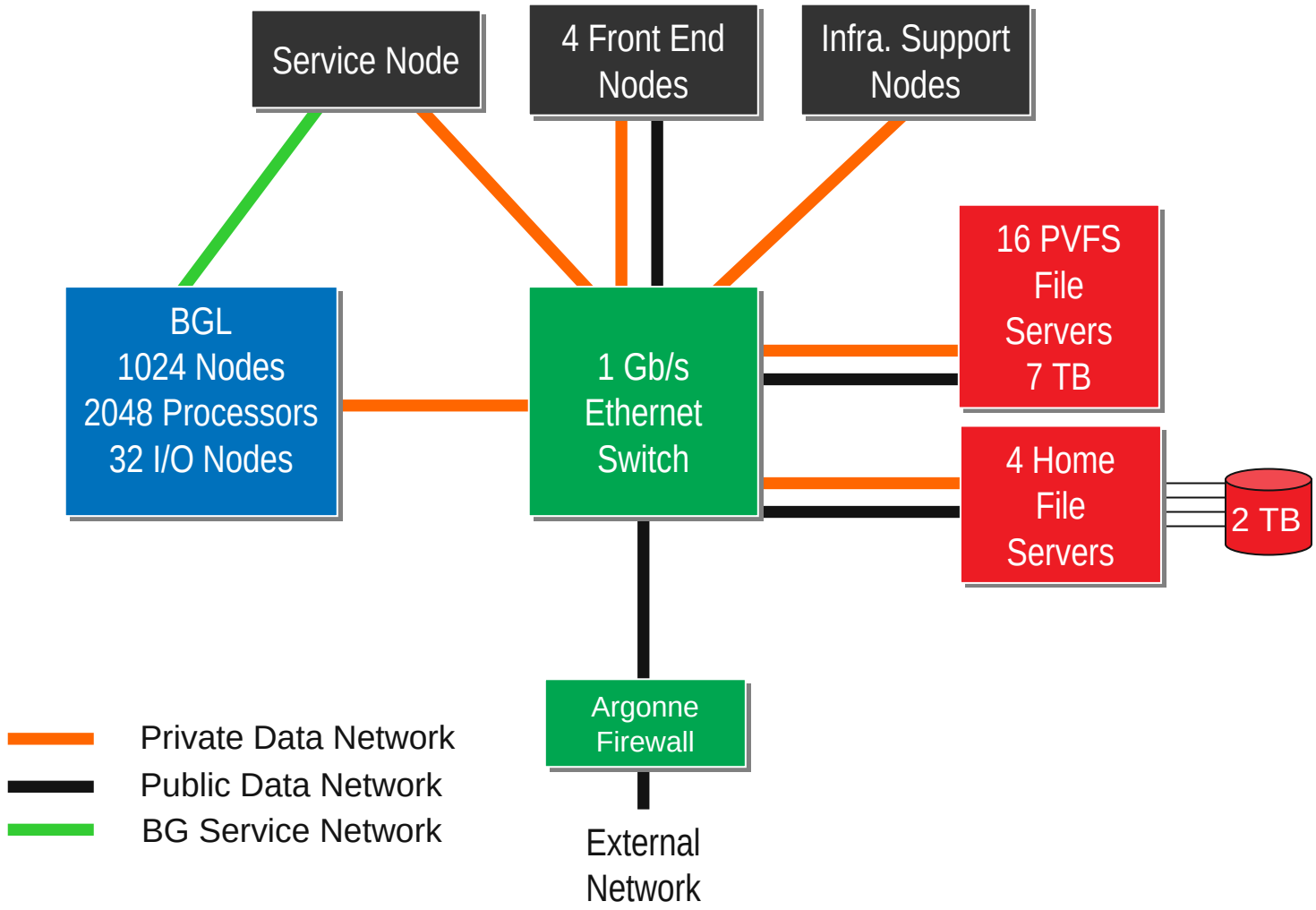


- Reaches high packaging density.
- Low system power requirements.
- Low cost per flops.

Record 280TF Linpack benchmark on 64K node BG/L at LLNL

Source: IBM

Argonne BGL System Architecture

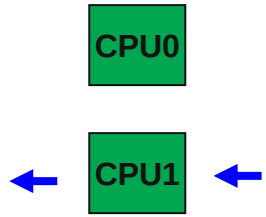


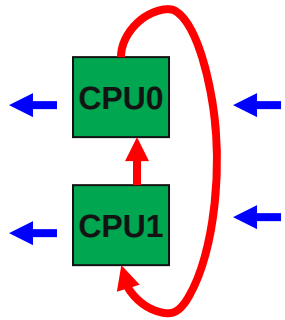
Programming Environment

- Fortran, C, C++ with MPI
- Linux: User accesses system through Front End nodes for compilation, job submission, debugging
- Compute Node OS: very small, selected services, I/O forwarding
- No OpenMP, no Threads
- Space sharing - one parallel job (user) per partition of machine, one process per processor of compute node
- Single executable image is replicated on each node
- Virtual memory limited to physical memory
- Libraries are statically linked

Applications Developer's View of Blue Gene

- Two CPU cores per node at 700 MHz
  - Each CPU can do 2 Float multiply-adds per cycle.
- Mode 1 (Co-processor mode - CO)
  - CPU0 does all the computations (512MB memory).
  - CPU1 does the communications.
  - Communications overlap with computation.
  - Peak compute performance is 5.6/2 = 2.8 GFlops.
- Mode 2 (Virtual node mode - VN)
  - CPU0, CPU1 independent "virtual tasks" (256MB each).
  - Each does own computation and communication.
  - The two CPUs talk via memory buffers.
  - Computation and communication cannot overlap.
  - Peak compute performance is 5.6 GFlops.
- 3-D Torus network with virtual cut-through routing
  - (point to point: MPI_ISEND, MPI_IRECV)
- Global combine/broadcast tree network
  - (collectives: MPI_GATHER, MPI_SCATTER)





BlueGene in the HPC Ecosystem

- **BlueGene is the most successful new architecture transition in recent HPC history.**
  - BG/L smashes performance records (LINPACK), sweeps major HPC awards (Gordon Bell), and leads in unexpected ways (HPCbench).
- **Efforts of LLNL, Argonne, and IBM were central in engaging the community early.**
  - Argonne and IBM formed the BlueGene Consortium, now with 67 member institutions.
  - LLNL hosted a valuable series of applications teleconferences.
  - Argonne reached out to a worldwide community with workshops in the USA, Europe, and Japan.
- **Argonne's BlueGene/L system provides broad community access.**
  - Open evaluation system has over 300 users; many codes ported.
  - DOE brought Argonne into INCITE program, and IBM joined forces with us.
- **The rich BlueGene community is active, informed, diverse, and interactive.**